Qualidade do Vinho Sob a Lente da Análise Exploratória de Dados

1st Gabriel Vasconcelos Fruet
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Ceará
gabrielfruet@alu.ufc.br

2nd Kelvin Leandro Martins
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Ceará
kelvinleandro@alu.ufc.br

3rd Mateus Pereira Santos
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Ceará
mateussantos14@alu.ufc.br

4th Pedro Leinos Falcão Cunha
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Ceará
pedrofalcao@alu.ufc.br

***Abstract*—Esse artigo se trata de uma análise exploratória do dataset *Wine Quality* com objetivo de encontrar as suas principais características, realizando analises univariadas, bivariadas e multivariadas. No artigo, também é realizado pré processamento dos dados, para análise dos componentes principais e para posteriormente ser possível utilizar os dados para regressão e classificação.**

***Index Terms*—Análise Exploratória, Análise de componentes principais,visualização de dados, vinhos**

I. INTRODUÇÃO

A qualidade do vinho é influenciada por diversas características físico-químicas, cuja análise sistemática pode auxiliar na otimização da produção. Este trabalho aplica técnicas de Análise Exploratória de Dados para identificar padrões e fatores determinantes da qualidade, utilizando um conjunto de dados público. Os resultados demonstram o potencial da Análise Exploratória de Dados como ferramenta inicial e essencial para compreender a relação entre os dados, diferenciá-los e embasar as decisões a serem tomadas após essa análise.

II. ANÁLISE EXPLORATÓRIA

A. Fundamentação Teórica

Inicialmente, é importante entender a necessidade da realização de uma análise exploratória e quais são as métricas utilizadas em seus diferentes tipos de análises.

A análise univariada é a análise que utiliza métricas que dependem de uma mesma variável do conjunto de dados. As principais métricas para essa análise são a média, o desvio padrão e a *skewness*.

A média representa a tendência central dos dados e é dada pela fórmula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

O desvio padrão descreve a dispersão dos dados em relação à média, isto é, quanto maior o desvio padrão, mais dispersos os dados estão ao redor da média. O desvio padrão é dado pela fórmula:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

E a *skewness*, que diz respeito à simetria dos dados, dessa forma, quando ela se aproxima de zero, o conjunto de dados é mais simétrico. Quando o valor é maior que zero, os dados tendem a uma assimetria à direita da média e, quando menor que zero, os dados tendem a uma assimetria à esquerda da média. [1] A *skewness* é dada pela fórmula abaixo:

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3 \tag{3}$$

Estas métricas fornecem uma visão inicial sobre a distribuição dos dados e podem indicar a necessidade de transformações ou ajustes para facilitar análises subsequentes.

Além dessas métricas, a análise univariada pode ser complementada por representações gráficas, como histogramas e boxplots, que ajudam a visualizar a distribuição e identificar valores atípicos ou irregularidades nos dados. Essa etapa é particularmente relevante na análise da qualidade do vinho, uma vez que características como pH, acidez e teor alcoólico podem apresentar padrões específicos dependendo da classe (vinho tinto ou branco), como será demonstrado na análise condicional apresentada neste trabalho.

A análise bivariada é a análise das variáveis duas a duas, buscando entender como uma variável influencia na outra. A métrica que define o quanto uma variável influencia em outra pode ser expressa através da correlação. A quantidade de correlações em um conjunto de dados é expressa por:

$$n^o correlações = \binom{n^o features}{2} \tag{4}$$

Dessa forma, um conjunto de 11 *features* possui 55 correlações, sendo difícil entender como esses valores se comportam em uma tabela. Uma das formas de visualizar esses dados de melhor forma é através de um mapa de calor, permitindo ver todas as correlações em conjunto.

Por fim, para a realização de uma análise multivariada, é necessária a redução de dimensionalidade, sabendo que um conjunto de dados com 11 *features* não é representado graficamente. Dessa forma, para realizar essa análise, utiliza-se o método do PCA, que define novas componentes a partir de uma combinação linear das *features* do conjunto de dados, em que essas combinações lineares concentram a maior variância possível. [1]

Assim, a partir das componentes principais definidas através do PCA, é possível representar graficamente um conjunto de dados com n *features* minimizando a perda de informação.

A realização da análise exploratória é crucial para reduzir a dimensionalidade, ajustar os dados e selecionar as variáveis mais relevantes para a modelagem preditiva.

B. Descrição dos dados

O conjunto de dados *Wine Quality* [2] possui 6497 amostras e 12 variáveis (ou *features*). Os vinhos estão divididos em duas classes: vinho tinto e vinho branco, sendo 4898 do tipo vinho branco (75.39%) e 1599 do tipo vinho tinto (24.61%). As *features* do dataset são [3]:

- **Acidez Fixa (g/dm³)**: Concentração de ácido tartárico na amostra.
- **Acidez Volátil (g/dm³)**: Concentração de ácido acético na amostra.
- **Ácido Cítrico (g/dm³)**: Concentração de ácido cítrico na amostra.
- **Açúcar Residual (g/dm³)**: Quantidade de açúcar restante após o processo de fermentação.
- **Dióxido de Enxofre Livre (g/dm³)**: Concentração de $SO_{2(g)}$.
- **Dióxido de Enxofre Total (g/dm³)**: Indica a concentração total de $SO_{2(g)}$, incluindo a fração dissolvida.
- **Densidade (g/dm³)**: Densidade do vinho.
- **pH**: Potencial de hidrogênio, variando de 0 (ultra ácido) a 14 (hiper alcalino).
- **Sulfatos (g/dm³)**: Concentração de sulfato de potássio na amostra.
- **Álcool (% por volume)**: Concentração de etanol na amostra.

C. Análise univariada incondicional

A tabela I mostra a média, desvio padrão e *skewness* (assimetria) para cada *feature*. A grande maioria das variáveis possuem baixo desvio em torno de sua média, exceto açúcar residual e dióxido de enxofre total.

A figura 1 apresenta a distribuição das variáveis por meio de seus histogramas. A variável cloretos apresenta uma concentração maior no lado esquerdo, indicando uma grande assimetria, como evidenciado pelo valor de *skewness* encontrado na tabela I, onde foi o preditor com maior assimetria. Já o pH dos vinhos está predominantemente concentrado entre os valores de 3 e 4.

TABLE I
ESTATÍSTICAS DO DATASET

Feature	Média	Desvio Padrão	*Skewness*
fixed acidity	7.215	1.296	1.723
volatile acidity	0.340	0.165	1.495
citric acid	0.319	0.145	0.472
residual sugar	5.443	4.758	1.435
chlorides	0.056	0.035	5.400
free sulfur dioxide	30.525	1.749	1.220
total sulfur dioxide	116.745	56.522	-0.001
density	0.995	0.003	0.504
pH	3.219	0.161	0.387
sulphates	0.531	0.149	1.797
alcohol	10.492	1.193	0.566
quality	5.818	0.873	0.190

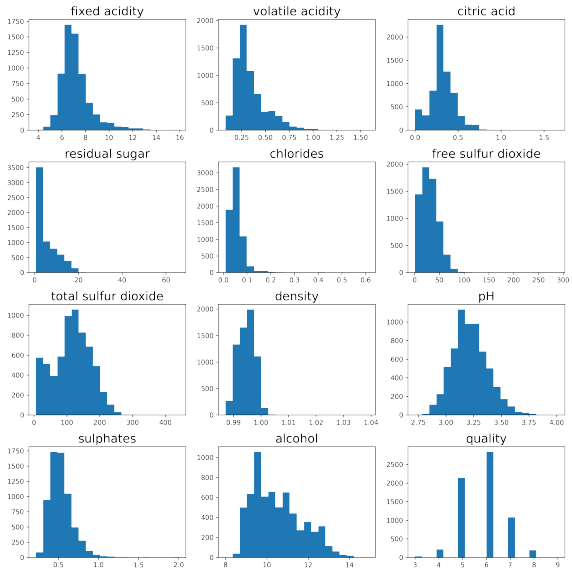

Fig. 1. Histograma de preditores.

A figura 2 ilustra o boxplot de algumas variáveis como qualidade e álcool. A maioria dos valores de qualidade se concentra entre aproximadamente 5 e 6. Há também alguns outliers acima de 7 e abaixo de 4, indicando que vinhos com tais valores não são comuns neste conjunto de dados. Em relação ao dióxido de enxofre total, a maioria dos valores está concentrada entre aproximadamente 80 e 150, porém há outliers significativos acima de 300, indicando amostras com altos níveis de dióxido de enxofre em comparação ao restante.

D. Análise univariada condicional à classe

As tabelas II e III mostram estatísticas para vinhos do tipo tinto e branco, respectivamente. Observa-se que, em geral, os vinhos brancos possuem uma média de qualidade ligeiramente superior (5.878) em comparação aos vinhos tintos (5.636). Além disso, os vinhos brancos tendem a ter menor variabilidade em características como acidez fixa (6.855 para

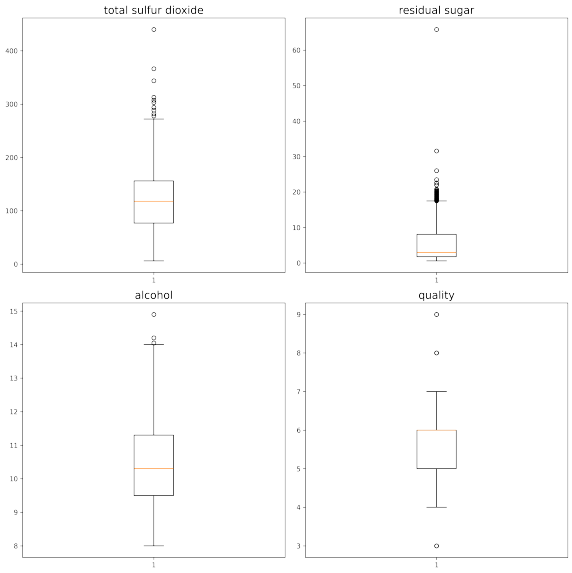

Fig. 2. Boxplot de preditores.

TABLE II
ESTATÍSTICAS DO VINHO TINTO

Feature	**Média**	**Desvio Padrão**	***Skewness***
fixed acidity	8.320	1.741	0.983
volatile acidity	0.528	0.179	0.672
citric acid	0.271	0.195	0.318
residual sugar	2.539	1.410	4.541
chlorides	0.087	0.047	5.680
free sulfur dioxide	15.875	10.460	1.251
total sulfur dioxide	46.468	32.895	1.516
density	0.997	0.002	0.071
pH	3.311	0.154	0.194
sulphates	0.658	0.170	2.429
alcohol	10.423	1.066	0.861
quality	5.636	0.808	0.218

TABLE III
ESTATÍSTICAS DO VINHO BRANCO

Feature	**Média**	**Desvio Padrão**	***Skewness***
fixed acidity	6.855	0.844	0.648
volatile acidity	0.278	0.101	1.577
citric acid	0.334	0.121	1.282
residual sugar	6.391	5.072	1.077
chlorides	0.046	0.022	5.023
free sulfur dioxide	35.308	17.007	1.407
total sulfur dioxide	138.361	42.498	0.391
density	0.994	0.003	0.978
pH	3.188	0.151	0.458
sulphates	0.490	0.114	0.977
alcohol	10.514	1.231	0.487
quality	5.878	0.886	0.156

branco contra 8.320 para tinto) e acidez volátil (0.278 para branco contra 0.528 para tinto).

A concentração de açúcar residual é significantemente maior nos vinhos brancos (6.391) do que nos tintos (2.539), cerca de 2.5 vezes maior, o que pode refletir diferenças no processo de vinificação ou no estilo de vinho. Além disso, a quantidade de dióxido de enxofre livre também é maior nos vinhos brancos (35.308 contra 15.875 nos tintos), sugerindo um uso mais intensivo de preservativos no vinho branco. A maior diferença entre as concentrações ocorreu no dióxido de enxofre total, 138.361 contra 46.468 nos tintos, uma razão de aproximadamente 3 vezes.

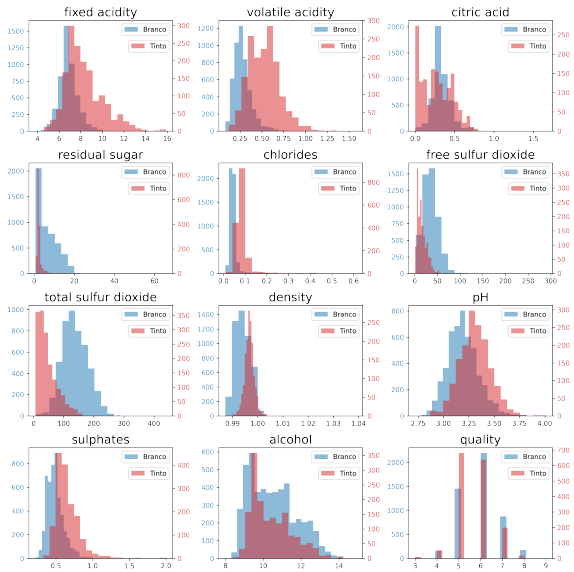

Fig. 3. Histograma de preditores condicionados à classe.

A figura 3 mostra a distribuição de cada variável condicionados as classe vinho (de vermelho) e branco (em azul). Na lateral esquerda de cada histograma há a contagem para os vinhos do tipo branco e o mesmo acontece para o vinho tinto na lateral direita.

Os vinhos tintos possuem valores de acidez volátil mais elevados, assim como sua distribuição é mais simétrica, possuindo *skewness* de 0.672, enquanto para vinhos brancos eles são mais concentrados na esquerda, com *skewness* de 1.577. De maneira semelhante em relação ao dióxido de enxofre total, os vinhos brancos possuem níveis totais mais elevados, enquanto os tintos apresentam valores predominantemente baixos, com uma distribuição mais assimétrica.

E. Análise bivariada

A análise bivariada, em complemento a univariada, é um estudo de como as *features* de cada classe vão se relacionar com as *features* da outra. Essa abordagem é essencial para identificar dependências, correlações e padrões que podem existir entre pares de variáveis no conjunto de dados. No contexto deste trabalho, a análise bivariada permitirá entender como as características físico-químicas dos vinhos interagem entre si e influenciam a qualidade final do produto.

Nesse sentido, a fim de evidenciar a correlação entre as variáveis de cada classe, foi utilizada a matriz de correlação correspondente a cada classe com o coeficiente de Pearson. O coeficiente de correlação de Pearson [4] ρ, mede o quão correlacionados dois dados são. Quanto maior o coeficiente de correlação, mais correlacionados são. Caso o coeficiente seja positivo, há uma correlação positiva (i.e o aumento de um está relacionado com o aumento do outro). Caso contrário, há uma correlação negativa (i.e o aumento de um está relacionado com a diminuição de outro). A formula para calcularmos o coeficiente de Pearson [4] entre duas *features* é a seguinte:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

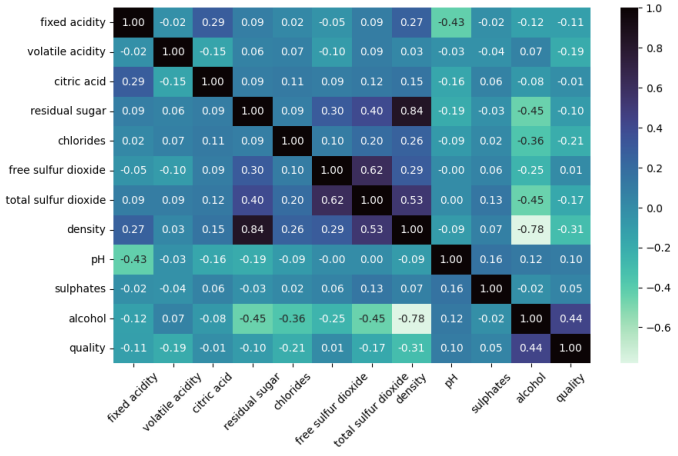

Fig. 4. Matriz de Correlação para o vinho branco.

Na figura 4 é evidente que há uma forte correlação positiva entre as variáveis *densidade* e *açúcar residual* e, além desta, uma forte correlação negativa entre as variáveis *álcool* e *densidade*, o que indica que a densidade é afetada, respectivamente, diretamente proporcional pelo açúcar cristalizado e inversamente proporcional pelo teor alcoólico.

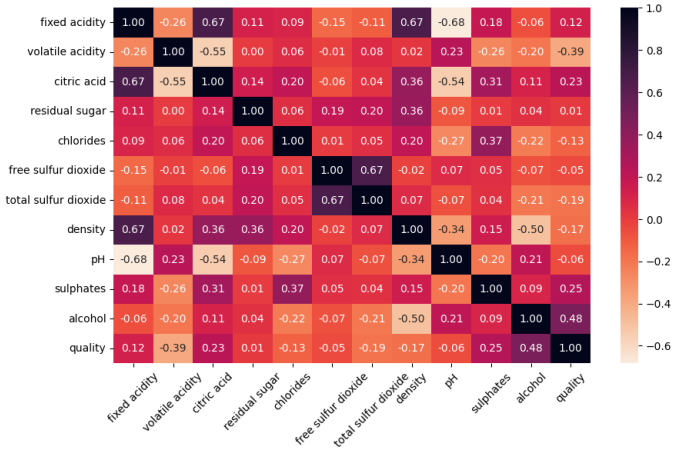

Fig. 5. Matriz de Correlação para o vinho tinto.

Já na figura 5, é claro que há uma forte correlação positiva entre os pares de variáveis *ácido cítrico* e *acidez fixa*, *dióxido de enxofre total* e *dióxido de enxofre livre*, *densidade* e *acidez fixa* e, além destas, há uma forte correlação negativa entre as variáveis *pH* e *acidez fixa*.

Devido a grande quantidade de *features* no dataset, acreditamos que basta plotarmos a análise bivariada entre dados que são correlacionados.

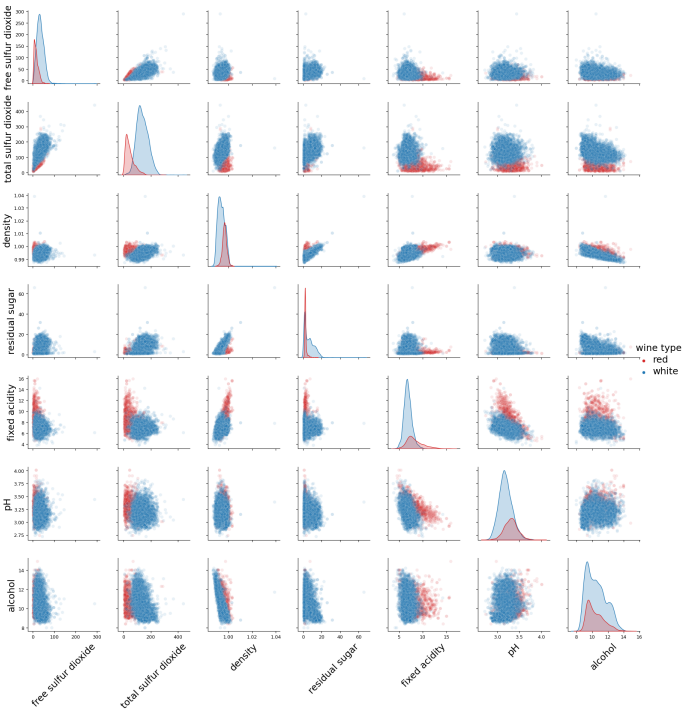

Fig. 6. Gráfico em pares da análise bivariada

Com base na fígura 6, podemos ver que os dados que citamos anteriormente, realmente possuem uma tendência de linearidade quanto plotados entre sí. Uma das colinearidade mais faceis de perceber é a da `density` por `fixed acidity`, que possui um coeficiente de correlação positiva de 0.67. Com base nisso, podemos ver que, quando `density` cresce, `fixed acidity` cresce também.

F. Analise das Componentes Principais

Para encontrar as componentes principais, realizamos uma decomposição em valores singulares da matriz de correlação $\boldsymbol{R_x}$.

$$\boldsymbol{R_x} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^H$$

A matriz $\boldsymbol{\Lambda}$ representa os autovalores da matriz $\boldsymbol{R_x}$, que são interpretados também como a quantidade de variação presente naquela direção.

Para encontrarmos o nosso dado $\bar{\boldsymbol{x}}(\boldsymbol{n})$ na base das componentes principais, basta aplicarmos a seguinte equação:

$$\bar{\boldsymbol{x}}(n) = \boldsymbol{U}\boldsymbol{V}^H\boldsymbol{x}(n)$$

Ao aplicarmos a analise das componentes principais, através da expansão de *Kahurne-Loève*, as variâncias características presentes no conjunto de dados foram concentradas em cada uma das componentes principais. Essa concentração de variâncias características possuem uma propriedade importante, que é a ortogonalidade, ou seja, cada direção de variância gerada pelo método das componentes principais é totalmente descorrelacionada de cada uma das outras direções.

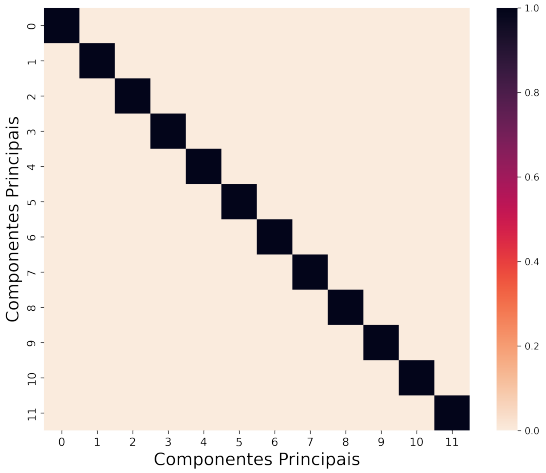

Fig. 7. Coeficiente de correlação de Pearson entre as componentes principais

Como podemos observar na figura 7, a correlação entre as componentes principais é sempre zero, já que elas formam uma base ortonormal, ou seja, não possuem nenhum tipo de colinearidade. As novas *features* que a aplicação do PCA gera, também podem ser chamadas de ”sinais brancos”, já que não possuem colinearidade.

Após extrairmos as componentes principais, podemos visualizar a quantidade de variância retida por cada uma delas através de um *scree plot* [1]. No escopo do nosso trabalho, escolhemos somente as duas primeiras componentes, devido a melhor capacidade visualização. Porém, em um cenário real, provavelmente deveriamos considerar escolher mais componentes, já que temos somente 47.4% dos dados explicados por elas.

Tendo em vista que temos 47.4% de variância retida nas duas primeiras componentes principais, podemos verificar o quão significativa é essa variância característica para classificar o tipo de vinho, já que o método PCA não leva em conta a classificação que buscamos, mas tão somente a forma dos dados e suas relações.

Como podemos ver, há uma divisória bem clara entre os tipos de vinho. Ou seja, as variâncias características extraídas pelo PCA foram bastante significativas para classificar os tipos de vinhos.

III. Resultados

A partir das analises realizadas, a princípio, podemos observar que é possível analisar os dados a partir de suas duas

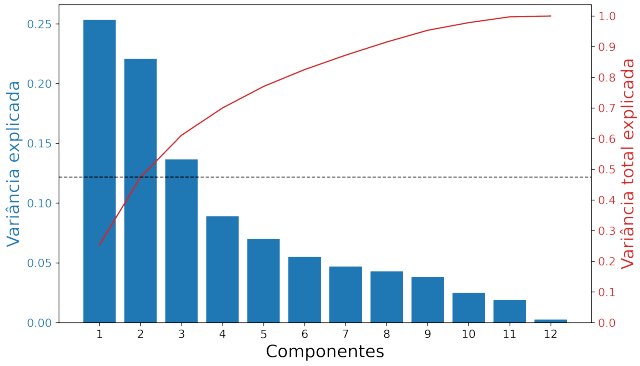

Fig. 8. *Scree plot* das componentes principais

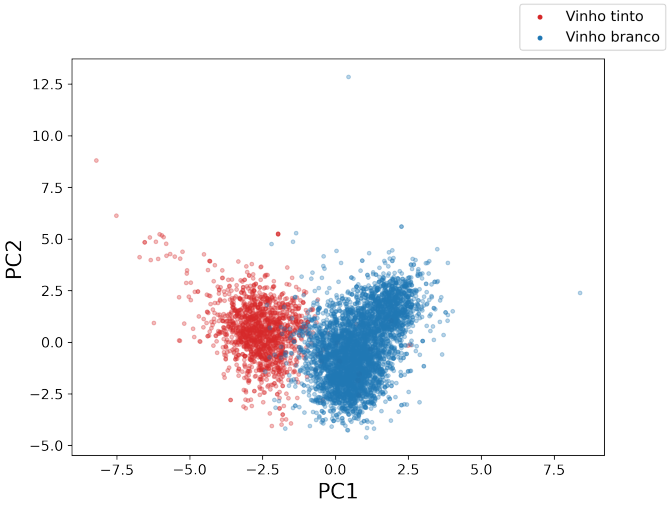

Fig. 9. Projeção do PCA em um espaço bidimensional

classes: Vinho Tinto e Vinho Branco, além disso, cada uma dessas classes possui características próprias. Isto é, apesar dos vinhos brancos e vinhos tintos possuírem *features* como acidez fixa e volátil, que não se diferem tanto, possuem outras como a quantidade de dióxido de enxofre livre com diferenças notáveis.

Essas diferenças para as distribuições das *features* com relação às classes ficam evidentes ao realizar a análise dos componentes principais (PCA), que destaca graficamente os agrupamentos de vinho tinto e vinho branco.

A análise exploratória de dados aplicada ao conjunto ”Wine Quality” permitiu identificar padrões relevantes e relações determinantes entre as *features* e a qualidade dos vinhos. A aplicação das análises univariada, bivariada e multivariada revelou que, embora algumas características sejam compartilhadas entre os vinhos tinto e branco, diferenças significativas foram observadas, como na concentração de dióxido de enxofre e no açúcar residual. Esses aspectos ressaltam a importância de uma abordagem exploratória detalhada para compreensão dos dados.

Além disso, o uso da análise de componentes principais (PCA), mostrou-se eficaz para reduzir a dimensionalidade do conjunto de dados e evidenciar agrupamentos claros entre os tipos de vinho, demonstrando o potencial de técnicas multivariadas para simplificar e visualizar a complexidade dos dados.

Portanto, os resultados destacam como a análise exploratória de dados fornece subsídios fundamentais para etapas subsequentes, como a modelagem preditiva, e contribui diretamente para o entendimento das variáveis mais relevantes que impactam a qualidade do vinho.

REFERENCES

[1] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. doi:10.1007/978-1-4614-6849-3

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. "Wine Quality," UCI Machine Learning Repository, 2009. [Online]. Available: https://doi.org/10.24432/C56S3T.

[3] P. Cortez, A. L. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," Decision Support Systems, vol. 47, pp. 547-553, 2009.

[4] Pearson, K. (1895). Notes on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242. doi:10.1098/rspl.1895.0041.

[5] Karhunen, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarum Fennicae. Series A. I. Mathematica-Physica*, vol. 37, pp. 1–79.

[6] Loève, M. (1948). Fonctions aléatoires de second ordre. *Revue Scientifique*, vol. 86, pp. 195–206.